

Text and Data Mining: One Year On

Dr Danny Kingsley and Dr Debbie Hansen
Office of Scholarly Communication
dak45@cam.ac.uk, dh554@cam.ac.uk

Outline of this meeting

Text and Data Mining

The past year

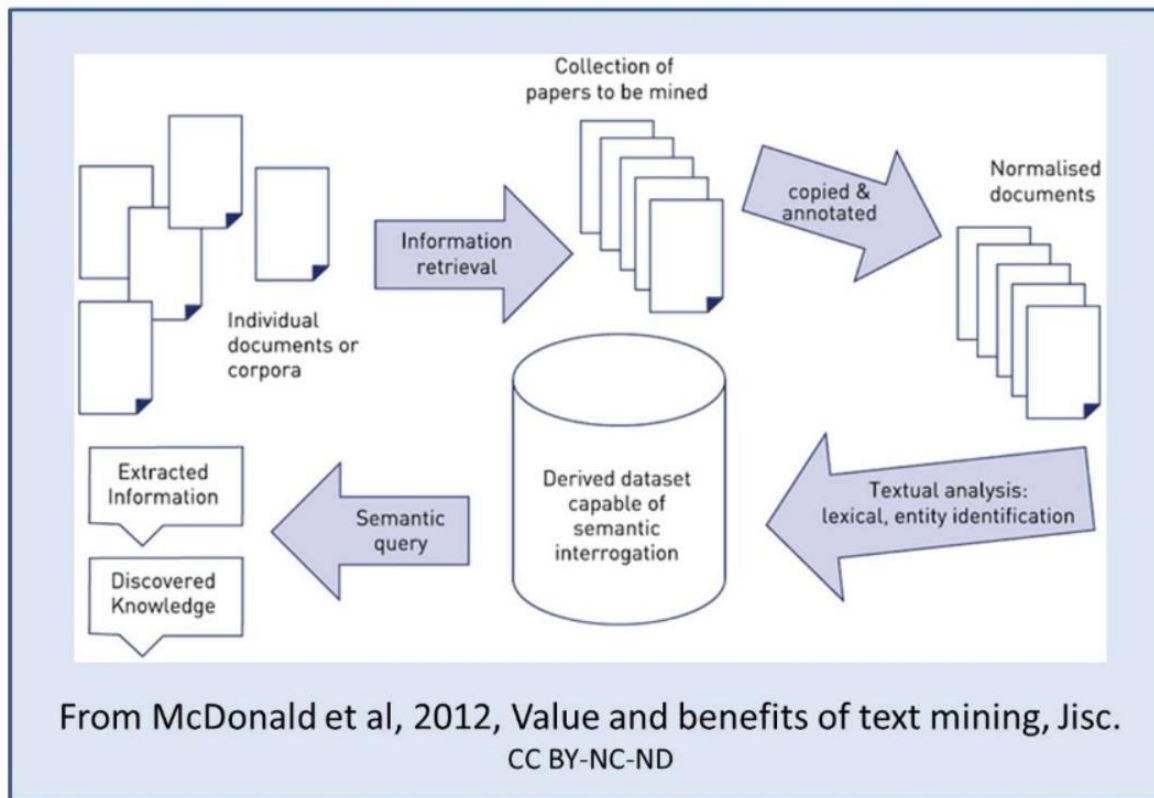
Currently in progress

Discussion

Towards developing a TDM library service

What is Text and Data Mining (TDM)?

TDM is the process of using digital techniques to explore large collections of machine-readable material, extracting new datasets and, by analysis, finding out new information about a topic



The past year

Text and Data Mining Services: What can Cambridge Libraries Offer? February 2017

~30 attendees, mainly library staff

Wide-ranging discussion and the following was agreed:

What could a Future Library TDM Support Service include?

- ACCESS to data from our own collections
- ADVICE on
 - Legal issues
 - What publishers allow
 - What data sets are available
 - What tools are available
- REGISTERS on
 - Data provided for mining
 - TDM projects
- FOSTER agreements with publishers

Next Steps

- MORE INFORMATION e.g.
 - Demonstrations – what TDM looks like
 - Case studies – that we can publish on-line
- WOULD BE USEFUL:
 - A Moodle group: TDM discussions/resources in one place
 - A TDM area on the library website
- WORKING GROUP: to help develop a University of Cambridge TDM service

(Kingsley, 2017, Notes from kick-off meeting 27 February)

The past year

TDM Moodle for communication within the community

The screenshot shows the Moodle interface for the Cambridge Text and Data Mining Community. The page has a blue header with the community name and a breadcrumb trail: Dashboard > My courses > University offices > University Library > Cambridge Text and Data Mining Community. On the left, there are navigation and administration links. The main content area features a central diagram titled 'TDM - how?' with a central blue box and surrounding labels: 'convert text', 'copying works', 'TDM tools', 'Apache', 'multiple copyright owners', 'structured data', 'computer processing', 'permissions', and 'time consuming'. Below this is a welcome message and a 'News' section with a link to 'Notes from kick-off meeting 27 February'. The 'Discussion' section includes a prompt to discuss text and data mining questions. The 'General resource list' provides links to various resources like 'Websites', 'Support for scholarly communication', 'JSTOR Analyser Tool', 'Slides from TDM workshop', 'OpenMinTeD', 'Text mining resources for the life sciences', and 'Slides from Jisc TDM Meeting'. On the right, there are sidebars for 'Course search', 'Latest announcements' (listing recent updates and news), 'Recent activity' (showing activity since Monday, 26 March 2018), 'Upcoming events' (stating there are none), and a 'Calendar' for March 2018.

<https://www.vle.cam.ac.uk/course/view.php?id=134922>

The past year

Developing a Research Library position statement on Text and Data Mining in the UK

Workshop at RLUK, 2017

Dr Danny Kingsley, Dr Debbie Hansen – University of Cambridge

Anna Vernon – Jisc

British Library 9th March 2017

- 19 mainly librarians from large/small/other Universities/research institutions; couple publishers.
- Discussion 1: Talk about any experiences you have had with TDM
- Discussion 2: Why your organisation is not actively supporting TDM or if it is, what are the top challenges you face?
- Discussion 3: If we were to draft a statement for a Service Level Agreement for publishers to assure us that if the activity is legal we will be reinstated within 1 hour (or something like that), what are the issues if we did this?

The past year

Developing a Research Library position statement on Text and Data Mining in the UK

Discussion 1: Talk about any experiences you have had with TDM

Following this discussion it was agreed that there was a need for:

- UNDERSTANDING regarding licensing
- A MECHANISM for advice - where to go within an institution and a publisher
- POLICY - development of procedures for handling TDM related requests
- ADDRESSING researcher behaviour – academics are not always concerned by copyright

The past year

Developing a Research Library position statement on Text and Data Mining in the UK

Discussion 2: Why your organisation is not actively supporting TDM? If it is, what are the top challenges you face?

Why not supporting?

- Practical reasons
 - challenges of handling physical media
 - risk of lockout
- Lack of demand
 - Not getting enquiries
 - Not much call
- Who is responsible?
 - Policy needed – issues not raised at academic level
 - Library service – how to scale up from individual queries?
 - Not joined up as organisations

Challenges:

- Understanding at the content-owner level
 - Library ensure people know their responsibilities
 - E.g. not for commercial use
 - Even if intrusive on the research process
- Time
 - Example of TDM contract between publisher and researchers
 - Took 2 years to finalise

The past year

Developing a Research Library position statement on Text and Data Mining in the UK

Discussion 3: If we were to draft a statement for a Service Level Agreement for publishers to assure us that if the activity is legal we will be reinstated within 1 hour (or something like that), what are the issues if we did this?

AGREED EXPECTATIONS

- Don't cut us off! Have a conversation first (and if you want to cut us off - prove there are all these activities happening in the UK)
- If you do cut us off and it turns out to be legitimate then we expect compensation for the time we were cut off
- Mechanisms for TDM where certain behaviours are expected - built into separate licensing agreements for TDM

The past year

Text and Data Mining LibGuide

The screenshot shows the 'Text & Data Mining: Home' page of the University of Cambridge LibGuide. The header includes the University of Cambridge logo and navigation links: Cambridge Libraries, Cambridge LibGuides, Subject Guides, Referencing Guides, and Research Skills+. The main heading is 'Resources'. Below it, a breadcrumb trail reads 'All libraries / LibGuides / Resources / Text & Data Mining / Home'. A search bar is located on the right. A navigation menu includes 'Home', 'What is TDM?', 'Support for your TDM', 'Law on TDM', 'Resources', and 'Cambridge TDM'. The main content area welcomes users and provides a brief description of the LibGuide. It features three columns of content: 'Contents' with links to 'What is TDM?', 'Support for your TDM', 'Law on TDM', 'Resources', and 'Cambridge TDM projects'; 'Contact us' with a contact form and a red envelope icon; 'OpenMinted Platform' with a description of the platform; 'New "TDM Test Kitchen"' with a description of the service and a link to 'Read more about the TDM Test Kitchen'; 'Next steps for TDM: Cambridge Symposium' with a link to 'Sharing thinking on TDM in Cambridge'; and 'New bite-size educational videos on TDM' with a list of videos. A Twitter feed at the bottom shows a tweet from Laura Jeffrey (@LauraJeffreyLib) about learning from @osctdm and @theUL.

UNIVERSITY OF CAMBRIDGE

Cambridge Libraries | Cambridge LibGuides | Subject Guides | Referencing Guides | Research Skills+

Resources

All libraries / LibGuides / Resources / Text & Data Mining / Home

Text & Data Mining: Home

Search this Guide

Home | What is TDM? | Support for your TDM | Law on TDM | Resources | Cambridge TDM

Welcome to this LibGuide supporting TDM practitioners in Cambridge, students and researchers considering a project employing TDM, and librarians fielding enquiries about TDM from their library users.


This guide is a work in progress and as far as we are aware, the first LibGuide on TDM in the UK! We want to make this useful to you, so please email us with any suggestions or ideas - write to ejournals@lib.cam.ac.uk. Thank you.

Contents

- What is TDM?
- Support for your TDM
- Law on TDM
- Resources
- Cambridge TDM projects

Contact us

If you have any questions about this service or have any feedback for us, please let us know.



OpenMinted Platform

OpenMinted is working on a platform that will be a gateway to many types of language data, including tagsets, ontologies, publications and corpora. The platform will also offer services and functionalities that are useful for text and data mining, and allow miners to share their tools and build their own.

New "TDM Test Kitchen"

The TDM Test Kitchen is an experimental service supported by Cambridge Digital Humanities, Cambridge University Library and Cambridge University Press.


The TDM Test Kitchen aims to:

- Explore the application of TDM (Text and Data-Mining) methods to CUP and UL collections.
- Provide a 'live' learning environment where researchers, CUP and library staff involved either using TDM methods or developing TDM support services can learn more about TDM methods, share good practice and exchange knowledge about how to overcome challenges.
- Facilitate discussion between researchers, the UL and CUP about how to develop TDM methods and services in future.

[Read more about the TDM Test Kitchen](#)

Next steps for TDM: Cambridge Symposium

Sharing thinking on TDM in Cambridge: Links to description and social media on the [Cambridge Symposium on TDM](#)

 **Laura Jeffrey** @LauraJeffreyLib

[Follow](#)

Learned so much at #osctdm from accidental TM of the Genizah @theUL , to top tips for supporting researchers from @senorctulhu

9:42 PM - 12 Jul 2017

New bite-size educational videos on TDM

OpenMinted "sets out to create an open, service-oriented e-Infrastructure for Text and Data Mining (TDM) or scientific and scholarly content. Researchers can collaboratively create, discover, share and re-use knowledge from a wide range of text-based scientific related resources in a seamless way".

OpenMinted has a Knowledge Base comprising a range of materials including visualizations of the TDM workflows, textual guides, Webinars, and training videos showing methods applied in practice by experts in the field:

- Key concepts and areas in TDM explained - part 1
- Key concepts and areas in TDM explained - part 2: Knowledge representation
- Key concepts and areas in TDM explained - part 3: Recommenders and filtering
- Key concepts and areas in TDM explained - part 4: Semantic search
- Key concepts and areas in TDM explained - part 5: Knowledge discovery

Developed by eResources team

<http://libguides.cam.ac.uk/tdm>

The past year

University of Cambridge TDM Symposium

12th July 2017

OBJECTIVE: To provide as much information as possible to the attendees regarding TDM

WHO CAME?

- postgraduate students
- early career researchers
- senior researchers
- administrative staff
- librarians
- publishers

WHAT WAS THERE TO LEARN?

- Talks and Show and Tell sessions, e.g.
 - Keynote from Kiera McNeice (Future TDM)
 - TDM overview, what the barriers are
 - TDM in action
 - E.g. TDM and the Cairo Genizah
 - Tools e.g. ChemDataExtractor, ContentMine
 - How a publisher supports TDM (PLoS)
 - How librarians support TDM
- Discussion: The future of TDM in policy

The past year

University of Cambridge TDM Symposium

Discussion outcomes

Why not more TDM happening in UK?

- Legal challenges
- Recruiting and training staff
- Lack of institutional TDM policies
- Lack of institutional/governmental leadership
- Publisher issues
 - What happens when a publisher cuts off access
 - Convoluted process to reinstate access
 - Impractical to inform publishers when TDM to take place
 - Threats from publishers if download a lot of material
 - Will TDM drive price increases?
- Lack of understanding of how the different TDM stakeholders work

NEXT STEPS

- Hands-on sessions
 - Discipline specific?
- Stakeholder communication
 - Publishers
 - Librarians
 - Researchers
 - Legal expertise

The past year

Work with FutureTDM/OpenMinTed/LERU

Involvement with EU TDM developments

INCLUDES:

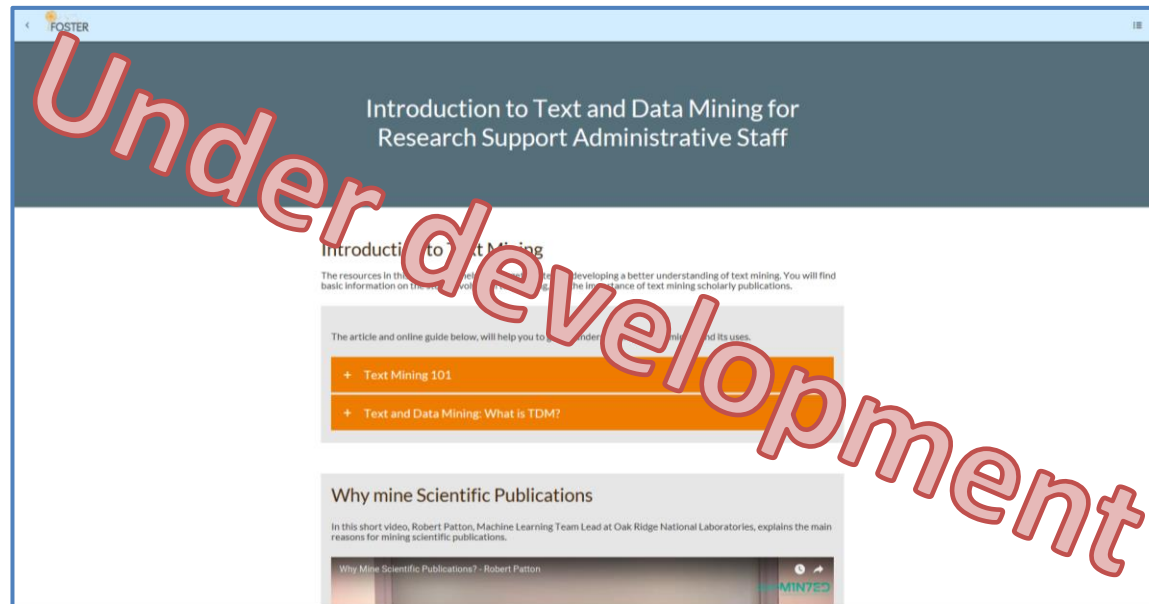
- Review of TDM Guidelines: outputs from the Future TDM Project
 - <https://www.futuretdm.eu/practitioner-guidelines/>
- University of Cambridge Librarian support for EU TDM Copyright Exception
 - Letters to our MEPs on the European Parliament Committee on Legal Affairs and our national Government minister for science calling for TDM Exception in legislative proposals of European Parliament and Council of the European Union to be strengthened and widened
- Future TDM Symposium (at iDSC '17, June 2017, Salzburg, Austria)
 - Panel discussion on Universities, TDM and the need for strategic thinking on educating researchers

Current activities in progress

OSC Working with OpenMinTeD: Text and Data Mining On-line Course

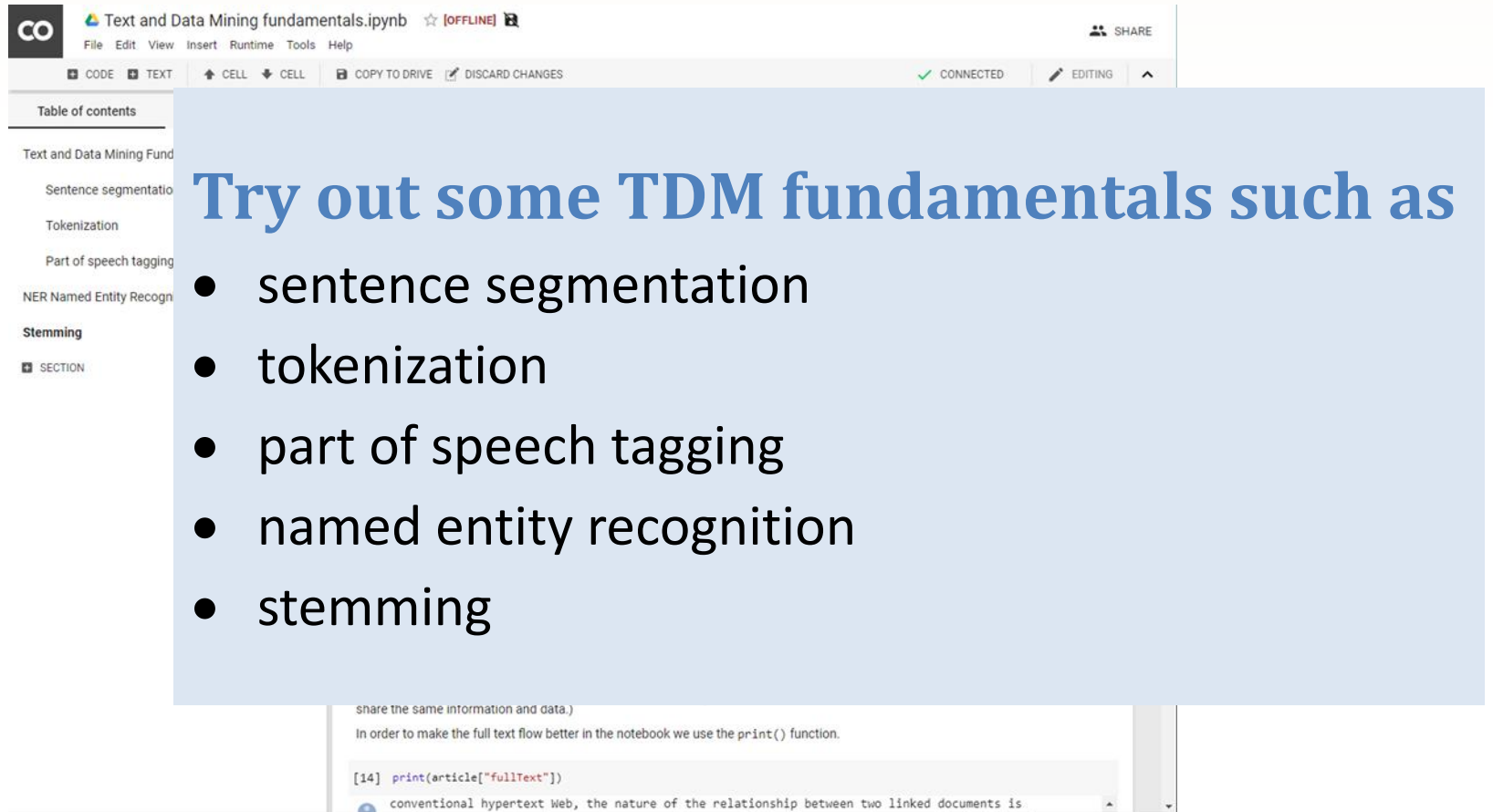
OpenMinTeD Foster on-line course

- Reading material
- Videos
- Quizzes
- Practical activities
- Glossaries



Current activities in progress

TDM on-line course: practical activities



The screenshot shows a Jupyter Notebook titled "Text and Data Mining fundamentals.ipynb". The interface includes a top menu bar with options like File, Edit, View, Insert, Runtime, Tools, and Help. Below the menu is a toolbar with icons for code, text, cell operations, and saving. A sidebar on the left displays a "Table of contents" with links to various sections: Text and Data Mining Fundamentals, Sentence segmentation, Tokenization, Part of speech tagging, NER Named Entity Recognition, Stemming, and a SECTION. The main content area features a large blue box with the text "Try out some TDM fundamentals such as" followed by a bulleted list: sentence segmentation, tokenization, part of speech tagging, named entity recognition, and stemming. Below this box, a code cell is visible, containing a print statement: `[14] print(article["fullText"])`. The notebook status bar at the bottom indicates it is "OFFLINE" and "CONNECTED".

Try out some TDM fundamentals such as

- sentence segmentation
- tokenization
- part of speech tagging
- named entity recognition
- stemming

Current activities in progress

TDM on-line course: practical activities

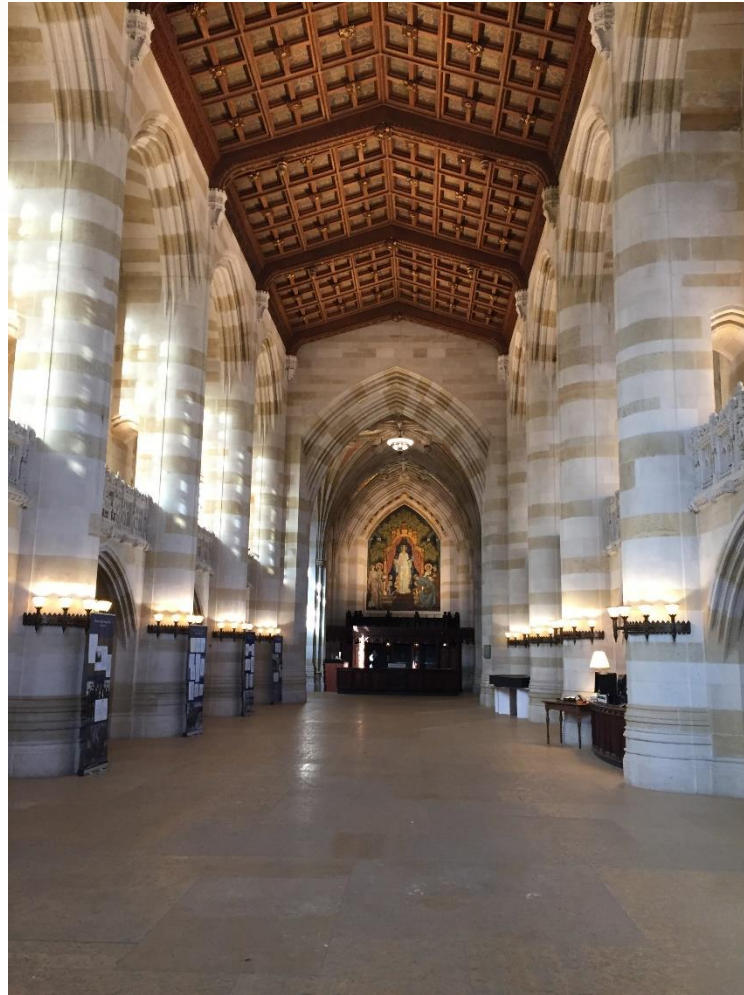
More advanced practical activity

- Mine selection of CORE articles
- Create a Word Cloud

The past year

Dr Danny Kingsley visit to Yale: January 2018

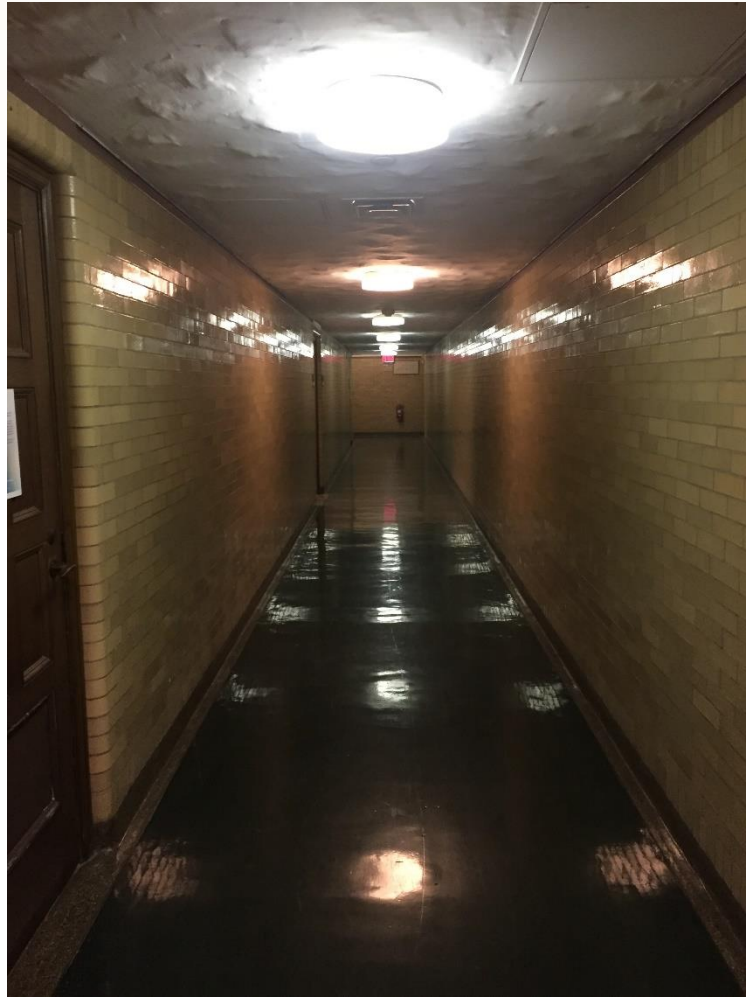
The hallowed halls of learning



Automated barriers to access stacks



Corridors in the Library



If you have a Yale card you can go through the barrier and are free to walk around here.

The DH Lab



Digital Humanities Lab



One Director, a UX person, an outreach person and two postdocs. Currently coming to end of four year grant, library is looking for \$ to get them on staff.

Moving to a bigger dedicated space on 1st floor in June with secure central glass box for sensitive data work

Digital Humanities Lab



These seats are for any student who wants to come in and work.
The red boxes contain data sets on hard drives that can be mined.



Digital Humanities Lab



An example of the events they run regularly

StatLab is part of the DH offering



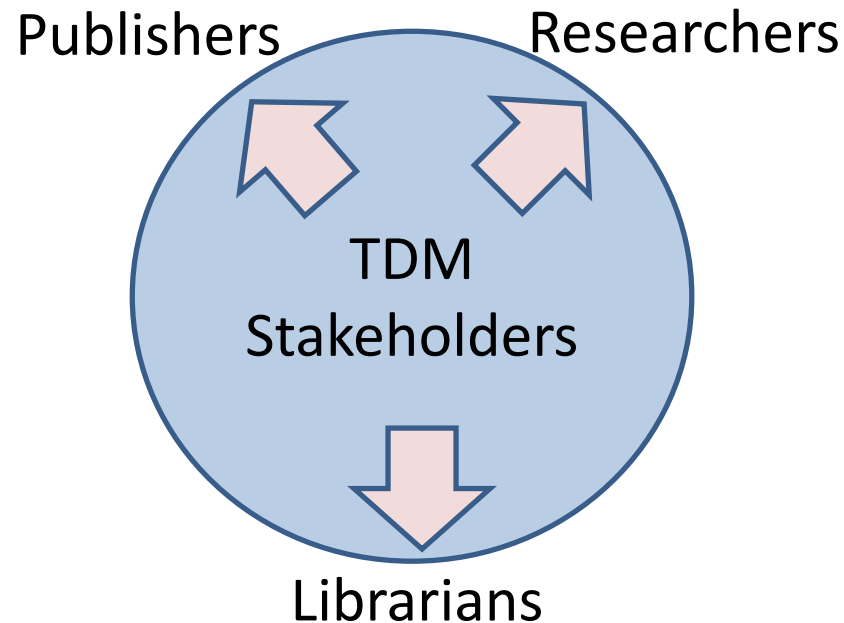
This IS the StatLab



A couple of postdocs are employed to man the desk

Current activities in progress

The TDM Test Kitchen Pilot Project



Current activities in progress

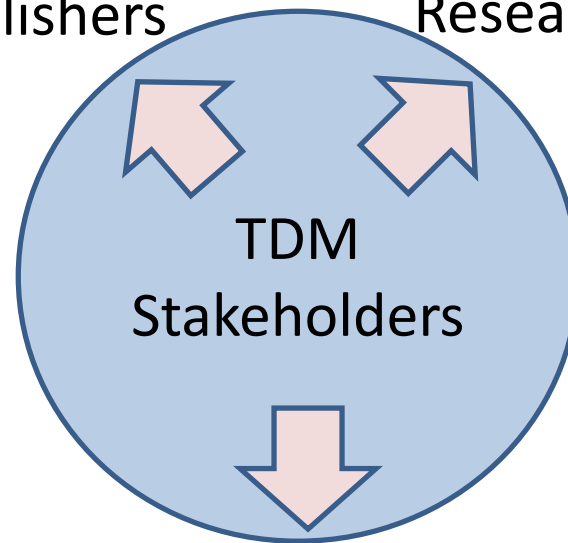
The TDM Test Kitchen Pilot Project

Cambridge University Press

Publishers

Cambridge Digital Humanities

Researchers



Librarians

University Library

Current activities in progress

The TDM Test Kitchen Pilot Project

<https://www.digitalhumanities.cam.ac.uk/Methods/tdm-testkitchen>

Experimental service aims to:

- Explore application of TDM methods to CUP and UL collections
- Provide a 'live' learning environment
 - Methods, developing support services, share good practice, overcoming challenges
- Facilitate discussion about how to develop TDM methods and services in the future

Current activities in progress

The TDM Test Kitchen Pilot Project

<https://www.digitalhumanities.cam.ac.uk/Methods/tdm-testkitchen>

Outline of Pilot Phase:

- Run from Jan – Jul 2018
- Open to Cambridge University graduate students and staff
- Advice provided to participants (from IP rights to data visualisation)
- Outputs – short report for website and presentation in Jun/Jul
- Create a set of case-studies which will inform future TDM service developments

Early lessons learned (from an observer)

Early queries and outcomes (UL):

- Investigation into a particular journal: some under subscription, some purchased, could get through a particular library?, obtained record of all payments to journal to date to check access rights, what about the pre-1950 content?.. (i.e. 1 journal, not straightforward)
- Is there a need to think about original digitization (OCR)?
- Where is the agreed corpus that is to be mined intended to be held?
- Do we need a dedicated Helpdesk? (TDM enquiry form on the TDM LibGuide at moment, access to this is eresources team)
- Requests made to publishers (**not just CUP**) for material for this pilot and agreement to supply copies on drives obtained :-
- It is emerging that there is often a cost associated with obtaining material for mining (O(£1000) per item) – from which budget can these costs be met?

Early lessons learned (from an observer)

Early queries and outcomes (publisher):

- Is the metadata and content in a form as required for the tools?
- Where is the agreed corpus that is to be mined intended to be held? A TDM-specific environment needed.
- Outputs for non-commercial research purposes and limited use of outputs from source material.
- Addendum to current licences.
- Content requested not always readily available on modern platforms which can hold up delivery time.

Early lessons learned (from an observer)

Early process (UL):

- Obtain and clarify request from researcher
- Investigate what available and rights and file format required
- Place request with publisher; publisher develops licence and terms and provides quote for costs
- Establish budget from which costs to be met; process payment and licence agreement if any
- Obtain material in agreed format
- Loan to researcher, who makes copy and returns original to UL
- Ensure researcher clear on terms of use.

Time for Discussion

Towards developing a TDM library service

Actions?

- Policy
- Procedures
- Costs
- Licensing
- Next pilot?